Galaxy Hands-on

Objectives

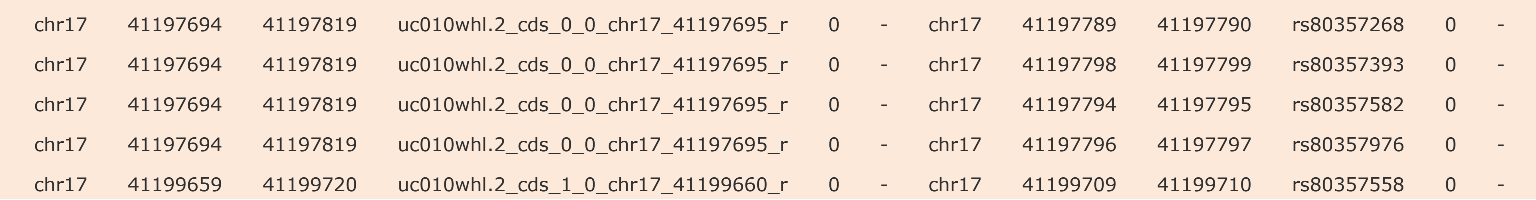
Visualize the top five BRCA1 exons with the higher number of SNPs, by designing a workflow using the three different methods available.

Learning Objectives

Navigate the Galaxy platform
Getting data from UCSC
Performing simple data manipulation
Understanding Galaxy's History system
Creating and editing workflows
Applying workflows to your data
History and workflow sharing
Brows shared items

Exercise

1. Setup working environment
   - a. Open the Galaxy website (URL: https://main.g2.bx.psu.edu)
   - b. Register through the User > Register Menu
   - c. Use the registered user name and password to login through the **User >Login menu**
2. Getting data from UCSC
   - a. Getting coding exons
     - i. On the tools panel click on **Get Data > UCSC Main**
     - ii. When you got the UCSC genome browser interface in the middle pane of Galaxy, than set region parameter by selecting '**position**' and enter '**chr17:41196312-41277500**' in the field. Leave all the other fields as default
     - iii. Click on get output
     - iv. In the new page '**output knownGene as BED**' select '**Coding Exons**' and click on send query to Galaxy
   - b. Getting SNPs
     - i. On the tools panel click on Get Data > UCSC Main
     - ii. This time select 'Variation and Repeats' under 'Group', and under track select 'All SNPs(137)'
     - iii. Finally in the region field put 'ch22' and click on the '**get output**'
     - iv. In the new page '**output snp137 as BED**' click on send query to Galaxy
   - c. Rename the imported files
     - i. In the history panel you can see your imported files
     - ii. Click on the small pen to edit the file features
     - iii. Rename the first file: Exons
     - iv. Rename the second file: SNPs
3. Finding the exons with the highest number of SNPs
   - a. Joining the exons with SNPs
     - i. Use '**Operate on Genomics Intervals > Join**' under the Tool panel
     - ii. You will get a third file where the 2 datasets are joined, and the data will look like (click on the the small eye):



chr17	41197694	41197819	uc010whl.2_cds_0_0_chr17_41197695_r	0	-	chr17	41197789	41197790	rs80357268	0	-
chr17	41197694	41197819	uc010whl.2_cds_0_0_chr17_41197695_r	0	-	chr17	41197798	41197799	rs80357393	0	-
chr17	41197694	41197819	uc010whl.2_cds_0_0_chr17_41197695_r	0	-	chr17	41197794	41197795	rs80357582	0	-
chr17	41197694	41197819	uc010whl.2_cds_0_0_chr17_41197695_r	0	-	chr17	41197796	41197797	rs80357976	0	-
chr17	41199659	41199720	uc010whl.2_cds_1_0_chr17_41199660_r	0	-	chr17	41199709	41199710	rs80357558	0	-

b. Count the number of the SNPs per exon, above we've seen that exon **uc010whl.2_cds_0_0_chr17_41197695_r** is repeated four times in the above dataset. Thus we can easily compute the number of SNPs per exon by simply counting the number of repetitions of name for each exon

i. Use the '**Join, Subtract, and Group > Group**' tool
ii. Choose the column 4 by selecting 'c4' in '**Group by column**'
iii. Click '**add new operation**'
iv. Under '**type**' select 'count', and select 'c4' under '**On column**'
v. Click on '**execute**'

c. Sorting exons by SNP count to find which exon has the highest number of SNPs

i. This is done with '**Filter and Sort > Sort**'
ii. Select the 4th file as a query
iii. Sort '**on column**' 'c2', make sure that the '**everything order**' option is set to 'Descending order'
iv. You can see the highest number of SNPs per exon is 67

d. Selecting top five, exons with the highest number of SNPs.

i. Use '**Text Manipulation > Select First**' tool
ii. Set the '**select first**' option to five (**5**)
iii. Click execute
iv. You have now a new file containing just the five first lines

e. Recovering exon info. And to know more we need to get back the positional information (coordinates) of these exons. This information was lost at the grouping step and now all we have is just two columns. To get coordinates back we will match the names of exons in dataset #6 (column 1) against names of the exons in the original dataset #1 (column 4)

i. Use '**Join, Subtract and Group > Compare two Queries**' tool
ii. '**compare**' 1:Exons file '**Using column**' c4 '**against**' 6:Select First on data 5 file, '**and column**' c1 '**To find**' (Matching rows of 1st query)
iii. Execute
iv. We got a seventh file in the history containing the top five exons with the highest number of SNPs with their coordinates information

f. Displaying data in genome browsers

i. Click on the file and you will see many options
ii. Click on the '**display at UCSC main**' option to see

4. Save workflow, extract and run workflow

a. Extract the workflow from the history by Click on '**Options**' on the top right of the Galaxy web page, select '**Extract Workflow**', name the workflow and then click on '**Create Workflow**'
b. Run the workflow by clicking on '**Workflow**' on top of the Galaxy web page, click on the workflow you created, and select 'Run'. Select input files and click on '**Run Workflow**'

5. Build a workflow from scratch to do the same analyses as step 1 through 7

i. Click on '**Workflow**' on top of the Galaxy web page, click on '**Create new Workflow**', name the workflow and click on '**Create**'
ii. Retake the same steps as the first workflow
iii. Add the inputs and rename the inputs
iv. Annotate the steps

v. Save the workflow by clicking on '**Options**' on upper right of Galaxy web page, select '**Save**'

6. Run the second workflows designed from the scratch
   a. Import data
      i. Create a new history
      ii. Go to the following link and import to the history this share dataset https://main.g2.bx.psu.edu/u/jeremy/h/galaxy-101-inputs
   b. Click on '**Workflow**' on top of the Galaxy web page, click on the workflow you created, and select '**Run**'. Select input files and click on '**Run Workflow**'

BED File format: .bed is tabular format developed for use with the UCSC genome browser, containing 3-12 columns of data, plus optional definition lines (comments).

The first three required BED fields are:

1. chrom - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).
2. chromStart - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. chromEnd - The ending position of the feature in the chromosome or scaffold. The chromEnd base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as chromStart=0, chromEnd=100, and span the bases numbered 0-99.

For more information about this file format:
http://genome.ucsc.edu/FAQ/FAQformat.html#format1